Local Discriminant Bases and Their Applications*

NAOKI SAITO
Schlumberger-Doll Research, Old Quarry Road, Ridgefield, CT 06877

RONALD R. COIFMAN
Department of Mathematics, Yale University, New Haven, CT 06520

Abstract. We describe an extension to the "best-basis" method to select an orthonormal basis suitable for signal/image classification problems from a large collection of orthonormal bases consisting of wavelet packets or local trigonometric bases. The original best-basis algorithm selects a basis minimizing entropy from such a "library of orthonormal bases" whereas the proposed algorithm selects a basis maximizing a certain discriminant measure (e.g., relative entropy) among classes. Once such a basis is selected, a small number of most significant coordinates (features) are fed into a traditional classifier such as Linear Discriminant Analysis (LDA) or Classification and Regression Tree (CARTTM). The performance of these statistical methods is enhanced since the proposed methods reduce the dimensionality of the problem at hand without losing important information for that problem. Here, the basis functions which are well-localized in the time-frequency plane are used as feature extractors. We applied our method to two signal classification problems and an image texture classification problem. These experiments show the superiority of our method over the direct application of these classifiers on the input signals. As a further application, we also describe a method to extract signal component from data consisting of signal and textured background.

Keywords: wavelet packets, local trigonometric transforms, feature extraction, classification, dimensionality reduction

1 Introduction

In analyzing and interpreting signals such as musical recordings, seismic signals, or stock market fluctuations, or images such as mammograms or satellite images, extracting relevant features from them is of vital importance. Often, the important features for signal analysis, such as edges, spikes, transients, or textures, are characterized by local information either in the time (or space) domain or in the frequency (or spatial frequency/wave number) domain or in both: for example, to discriminate seismic signals caused by nuclear explosions from the ones caused by natural earthquakes, the frequency characteristics of the primary waves, which arrive in a short and specific time window, may be a key factor; to distinguish benign and malignant tissues in mammograms, the sharpness of the edges of masses may be of critical importance.

In this paper, we explore how to extract relevant features from signals/images and discard irrelevant information for signal/image classification problems. In particular, we propose a fast algorithm to select an efficient basis (or coordinate system) from a large collection of orthonormal bases (consisting of wavelet packets and local trigonometric bases) to enhance the performance of a few classification schemes. This algorithm reduces the dimensionality of the problems by using these basis functions (which are well-localized in the time-frequency plane) as feature extractors. Since this basis illuminates the differences among classes, it can also be used to extract signal component from data consisting of signal and textured background.

The organization of this paper is as follows. In Section 2, we formulate the problem of feature extraction and classification and briefly review some pattern classification schemes used in our study. Then, in Section 3, we review the "best-basis paradigm" and a dictionary and a library of orthonormal bases which play a critical role for local feature extraction. Section 4 is a core material of this paper: we describe a fast algorithm for constructing a good local basis for classification problems. This is immediately followed by signal classification examples in Section 5 and an image texture classification problem in Section 6. In Section 7, we discuss a method of...
signal/"background" separation as a further application of such a basis.

We note that a concise version of this paper was announced earlier in [1] which also contains an algorithm for constructing a local basis for regression problems, and was presented in the SPIE conference [2]. The other aspects of our proposed method, including its applications to regression problems and examples using real datasets, can be found in [3–5].

2 Problem Formulation and Review of Pattern Classifiers

2.1 Formulation of a Signal Classification Problem

Let us first define appropriate spaces of input signals (or patterns), extracted features, outputs (or responses), and mapping functions among them. Let \( \mathcal{X} \subset \mathbb{R}^n \) denote a signal space (or a pattern space) which is a subset of the standard \( n \)-dimensional vector space and which contains all signals (or samples/patterns) under consideration. In this case, the dimensionality of the signal space is equivalent to the length of each signal. Let \( \mathcal{Y} = \{1, 2, \ldots, C\} \) be a set of the class or category names to which the input signals belong. We call this space a response space. Signal classification can be considered as a mapping function (usually many-to-one) \( d: \mathcal{X} \rightarrow \mathcal{Y} \) between these two spaces. Direct manipulation of signals in the signal space is prohibitive because: 1) the signal space normally has very high dimensionality (e.g., \( n \approx 1000 \) for a typical exploration seismic record per receiver, and for a typical CT scanner image, \( n = 512 \times 512 = 262,144 \)), and 2) the existence of noise or undesired components (whether random or not) in signals makes classification difficult. On the other hand, the signal space is overly redundant compared to the response space. Therefore, it is extremely important to reduce the dimensionality of the problem, i.e., extract only relevant features for the problem at hand and discard all irrelevant information. If we succeed in doing this, we can greatly improve classification performance both in its accuracy and efficiency. For this purpose, we set up a feature space \( \mathcal{F} \subset \mathbb{R}^k \) where \( k \leq n \) between the signal space and the response space. A feature extractor is defined as a map \( f: \mathcal{X} \rightarrow \mathcal{F} \), and a classifier (or predictor) as a map \( g: \mathcal{F} \rightarrow \mathcal{Y} \). Let \( \mathcal{F} = \{(x_i, y_i)\}_{i=1}^{N} \subset \mathcal{X} \times \mathcal{Y} \) be a training (or learning) dataset with \( N \) pairs of signals \( x_i \) and responses (class names) \( y_i \). This is the dataset to be used to construct a feature extractor \( f \). Let \( N_c \) be the number of signals belonging to class \( c \) so that we have \( N = N_1 + \cdots + N_C \). Also, let us denote a set of class \( c \) signals by \( \{x_i^{(c)}\}_{i=1}^{N_c} = \{x_i\}_{i \in I_c} \) where \( I_c \subset \{1, \ldots, N\} \) is a set of indices for class \( c \) signals in the training dataset with \( |I_c| = N_c \).

Preferably, the performance of the whole process should be measured by the misclassification rate using a test dataset \( \mathcal{T}' = \{(y'_i, x'_i)\}_{i=1}^{N'} \) which has not been used to construct the feature extractors and classifiers as \( (1/N') \sum_{i=1}^{N'} \delta(y'_i - d(x'_i)) \), where \( \delta(r \neq 0) = 1 \) and \( \delta(0) = 0 \). If we use the resubstitution (or apparent) error rates (i.e., the misclassification rates computed on the training dataset), we obviously have overly optimistic figures.

In this paper, we focus on the feature extractors of the form

\[
\hat{f} = \Theta^{(k)} \circ \Psi,
\]

where \( \Theta^{(k)}: \mathcal{X} \rightarrow \mathcal{F} \) represents the selection rule (e.g., picking most important \( k \) coordinates from \( n \) coordinates), and \( \Psi \in O(n) \), i.e., an \( n \)-dimensional orthogonal matrix. In particular, we consider matrices representing the orthonormal bases in the basis library (consisting of wavelet packets or local trigonometric bases) as candidates for \( \Psi \). As a classifier \( g \), we adopt Linear Discriminant Analysis (LDA) of R.A. Fisher [6] and Classification and Regression Trees (CART) [7].

In the following, we briefly review these two classification schemes. We note that other classifiers such as \( k \)-nearest neighbor (k-NN) [8], or artificial neural networks (ANN) [9] are all possible to use in our algorithm. The reader interested in comparisons of different classifiers is referred to the excellent review article of Ripley [9]. The useful information on pattern classifiers in general can be found in the books [10–13].

2.2 Linear Discriminant Analysis

Fisher’s LDA first tries to do its own feature extraction by a linear map \( A^T: \mathcal{X} \rightarrow \mathcal{F} \) (in this case not necessarily orthogonal matrix). This map \( A \) simultaneously minimizes the scatter of sample vectors (signals) within each class and maximizes the scatter of mean vectors of classes around the total mean vector. To be more precise, let \( m_c = (1/N_c) \sum_{i=1}^{N_c} x_i \) be a mean vector of class \( c \) signals. Then the total mean vector \( m \) can be defined as \( m = \sum_{c=1}^{C} \pi_c m_c \), where \( \pi_c \) is the prior probability of class \( c \) (which can be set to \( N_c/N \) without the knowledge on the true prior